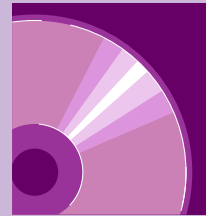


SIPP Public Use Files



This section covers basic concepts and topics that analysts need to understand when working with the SIPP public use files.

- *Types of SIPP Data Files*
- *Common Features Across SIPP Data Files*
 - Changes in Variable Names*
 - Survey Instrument Vs. Data Dictionary*
 - Identification/Description Variables*
 - Basic ID Variables*
 - Monthly Interview Status*
 - Identifying Persons*
 - Identifying Households*
 - Identifying Families*
 - Describing Relationships to Reference Persons*
 - Identifying Program Units*
 - Identifying Movers and Household Composition Changes*
 - Identifying States and Metro Areas*
 - Choosing Weights*
 - Income Topcoding*
 - Using Allocation Flags*

Types of SIPP Data Files

There are three types of public use files containing SIPP data: core wave files, topical module files, and full panel longitudinal research files.

Core Wave Files. Since 1990, these files have been issued in person-month format. They contain up to four records for each primary sample member and for each person who ever lived with a primary sample member during the reference period. Each record contains data from 1 of the 4 reference months in the wave.



Topical Module Files. For the 1996 Panel, these files contain one record for each person who was in the sample with a completed or imputed interview in the fourth month of the wave's reference period. Topical module files from previous panels contain one record for each person who was in the sample with a completed or imputed interview during the interview month (month 5), not the fourth month of the reference period.

Full Panel Longitudinal Research Files. These files are also referred to as "full panel files" and "longitudinal files." They contain one record for each primary sample member and for each person who ever lived with a primary sample person during the panel.

Common Features Across SIPP Data Files

The remainder of this section addresses features common to all three types of SIPP files. Although the features apply to each of the three file types, the files may differ in important ways with respect to the features. Those differences will be highlighted in subsequent sections of this tutorial.

Table 9-2 in the *SIPP Users' Guide* summarizes some of the file similarities and differences by topic.

Changes in Variable Names

For the 1996 Panel, most variable names changed from those used in previous panels. When appropriate, the *SIPP Users' Guide* presents both sets of names.

The technical documentation that users receive with their data files will include an item booklet for the 1996 Panel and the paper survey instrument for earlier panels. **tip**

The Survey Instrument and the Data Dictionary

With each order of a public use data file from the Census Bureau, users receive a set of technical documentation that includes, among other items, the survey instrument (or documentation of instrument screens and program code in the 1996 Panel) and a data dictionary.

Survey Instrument. The survey instrument is vital to understanding:

- What questions were asked
- How the questions were asked
- The order in which the questions were asked
- To whom the questions were asked
- The way in which the answers were recorded

SIPP *tip*

Appendix A of the *SIPP Users' Guide* contains a crosswalk of variable names for the 1993 and 1996 core wave files. Link to a view of Appendix A.

Section 1 - LABOR FORCE AND RECIPIENCY

(SHOW FLASHCARD J)

1. During the 4-month period outlined on this calendar, that is, from 4 months ago through last month, did ... have a job or business, either full time or part time, even for only a few days? Mark "Yes" for active duty in the Armed Forces, any temporary or part-time work, and work without pay in a family business or farm.

1000 ☐ Yes - Mark "Worked" (code 170) on ISS and SKIP to 4
1001 ☐ No

2a. Even though ... did not have a job during this period, did ... spend any time looking for work or on layoff from a job?

1002 ☐ Yes
1003 ☐ No - SKIP to 3a

b. Please look at the calendar. In which weeks was ... looking for work or on layoff from a job? Please answer by giving the week number that appears to the right of each week on the calendar. Mark (X) all that apply.

1004	<input type="checkbox"/> ALL	1009	<input type="checkbox"/> 7	1010	<input type="checkbox"/> 13
1005	<input type="checkbox"/> 1	1010	<input type="checkbox"/> 8	1011	<input type="checkbox"/> 14
1006	<input type="checkbox"/> 2	1011	<input type="checkbox"/> 9	1012	<input type="checkbox"/> 15
1007	<input type="checkbox"/> 3	1012	<input type="checkbox"/> 10	1013	<input type="checkbox"/> 16
1008	<input type="checkbox"/> 4	1013	<input type="checkbox"/> 11	1014	<input type="checkbox"/> 17
1009	<input type="checkbox"/> 5	1014	<input type="checkbox"/> 12	1015	<input type="checkbox"/> 18
1010	<input type="checkbox"/> 6				

c. Could ... have taken a job during any of those weeks if one had been offered?

1042 ☐ Yes - SKIP to 3a
1043 ☐ No

d. What was the main reason ... could not take a job during those weeks? Mark (X) only one.

1044 ☐ Already had a job
☐ Temporary illness
☐ School
☐ Other - Specify _____

3a. Even though ... did not have a job during this period, did ... do any work at all that earned some money?

1040 ☐ Yes - Mark "SS" on ISS
1041 ☐ No - SKIP to 3a, page 4

b. In which of the months shown on this calendar did ... do that work? Mark (X) all that apply.

1048 ☐ Last month
1049 ☐ 2 months ago
1050 ☐ 3 months ago
1051 ☐ 4 months ago

4. Did ... have a job or business, either full or part time, during EACH of the weeks in this period? Note that the person did not have to work each week.

1052 ☐ Yes
1053 ☐ No - SKIP to 3a

5a. Was ... absent without pay from ...'s job or business for any FULL weeks during the 4-month period?

1054 ☐ Yes
1055 ☐ No - SKIP to 5a, page 4

b. Please look at the calendar. In which weeks was ... absent without pay? Please answer by giving the week number that appears to the right of each week on the calendar. Mark (X) all that apply.

1056	<input type="checkbox"/> ALL	1061	<input type="checkbox"/> 7	1062	<input type="checkbox"/> 13
1057	<input type="checkbox"/> 1	1062	<input type="checkbox"/> 8	1063	<input type="checkbox"/> 14
1058	<input type="checkbox"/> 2	1063	<input type="checkbox"/> 9	1064	<input type="checkbox"/> 15
1059	<input type="checkbox"/> 3	1064	<input type="checkbox"/> 10	1065	<input type="checkbox"/> 16
1060	<input type="checkbox"/> 4	1065	<input type="checkbox"/> 11	1066	<input type="checkbox"/> 17
1061	<input type="checkbox"/> 5	1066	<input type="checkbox"/> 12	1067	<input type="checkbox"/> 18
1062	<input type="checkbox"/> 6				

c. What was the main reason ... was absent without pay from ...'s job or business during those weeks? Mark (X) only one.

1068 ☐ On layoff
☐ Own illness
☐ On vacation
☐ Bad weather
☐ Labor dispute
☐ New job to begin within 30 days
☐ Other - Specify _____

NOTES

Data Dictionary. The data dictionary describes four aspects of each variable:

- Definition
- Sample universe for the corresponding survey question
- Ranges for all legal values
- Location in the file

It is important that users understand that the data dictionary does not replicate the survey instrument. Analysts should therefore be aware of the following:

- Variables on the data files do not have a one-to-one correspondence with questionnaire items.
- The range of possible values of variables on the data files does not always correspond exactly with the response categories in the survey instrument or the data dictionary.
- Variable names in the data dictionary may not readily reflect the variable's content.
- Skip patterns will not be obvious from simply looking at the data dictionary. *tip*

Identification/Description Variables

Basic ID Variables in SIPP

The capacity to identify units across files allows SIPP users to:

- Follow participants over time
- Determine when an individual is present in the sample
- Verify the make-up of families and households

```

SURVEY OF INCOME AND PROGRAM PARTICIPATION,
1996 PANEL WAVE 1 TOPICAL MODULE DATA DICTIONARY

DATA      SIZE  BEGIN
D  SSUSEQ    5    1
T  SU: Sequence Number of Sample Unit - Primary
    Sort Key
U  All persons
V    1:50000 .Sequence Number

D  SSUID     12    6
T  SU: Sample Unit Identifier
    Sample Unit identifier This identifier is
    created by scrambling together the PSU,
    Segment, Serial, Serial Suffix of the
    original sample address. It may be used
    in matching sample units from different
    waves.
U  All persons
V  000000000000:999999999999 .Scrambled Id

D  SPANEL    4    18
T  SU: Sample Code - Indicated Panel Year
U  All persons
V

```

SIPP *tip*

Analysts should become familiar with the survey instrument before using the data. This will prevent confusion and help avoid problems. It is also helpful to refer to the survey instrument and data dictionary while working with the data.

The four most basic identification (ID) variables in SIPP include the following:

Sample Unit IDs. These uniquely identify each physical dwelling unit in the sample. The sample unit ID assigned to a person never changes. All people who have ever lived with a member of a given original sample unit share the same sample unit ID.

Current Address IDs. These identify the housing units occupied by one or more original sample members in a given month. They are assigned within sample units.

Entry Address IDs. These are the current address IDs for each sample member's initial address. They do not change when a person moves.


Person Number IDs. Person numbers are assigned sequentially, within each wave and each household, to all primary and secondary sample members when they first enter the sample.

These four variables have different names in the different types of public use files. [Link to a table that includes the names of the ID variables in the three types of files.](#)

Monthly Interview Status

The monthly interview status variable, which has values of 0, 1, or 2, helps analysts determine whether or not to use the data for a person in a given month.

Analysts should use data only for those months in which a person's interview status is equal to 1. Examining either the weight variable or the variable used in the analysis itself, as is often done with other data sources, will lead the SIPP user astray. See Chapter 9 of the *SIPP Users' Guide* for more information.

Analysts should ignore any data for months in which a person's interview status is coded either 0 (indicating a person was not in the sample that month) or 2 (indicating a noninterview for that month). 

SIPP *tip*

Because the person-month core wave files and the 1996 topical module files contain records only for those months that a person has an interview status code of 1, the monthly interview status variables in those files can be safely ignored.

Identifying Persons

Analysts may need to identify which records belong to which individual in SIPP data files. For example, analysts may need that information to combine data from file types, to link family members, and to identify the recipient of government transfer income.

Each person in SIPP can be identified by the combination of sample unit ID, entry address ID, and person number. *tip*

Identifying Households

A household consists of all people who occupy a housing unit, regardless of their relationships to one another. The many variations of households include, for example:

- A group of friends sharing a townhouse
- A single person in an apartment
- A family in a house

Each household contains one household reference person—the owner or renter of record.

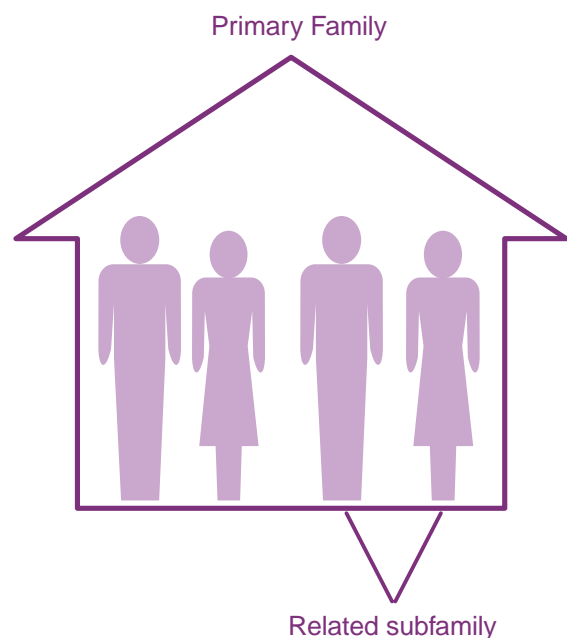
Identifying Families

The Census Bureau defines a family as a group of two or more people who reside together and are related by birth, marriage, or adoption. There are several types of families that the Census Bureau distinguishes:

- A primary family contains the household reference person and all of his or her relatives.
- A related subfamily is a family unit within the primary family whose members are related to, but do not include, the household reference person. An example would be a son and his wife living with the son's parents, one of whom is the household reference person.

SIPP *tip*

For the 1996 Panel, analysts do not need to use the entry address to uniquely identify individuals.



- An unrelated subfamily, or secondary family, is a family living in the household whose members are not related to the household reference person.
- A primary individual is a household reference person who lives alone or with nonrelatives. The Census Bureau sometimes treats primary individuals as one-person families and refers to them as pseudo-families.
- A secondary individual is not a household reference person and is not related to other people in the household. The Census Bureau also sometimes refers to such individuals as pseudo-families.

The Census Bureau has two principal methods for distinguishing families:

- The first method defines a family as all persons who are related and living together.
- The second method is similar to the first but excludes members of related subfamilies.

The variables and numbering schemes associated with these two methods allow analysts to construct various family units, including multigenerational families.

The various types of data files in SIPP, however, contain different identification information about family relationships. In fact, the topical module files contain no information for directly identifying different types of families. Thus, the analytic tasks for establishing family membership vary across file types. These differences will be highlighted in subsequent sections of the tutorial.

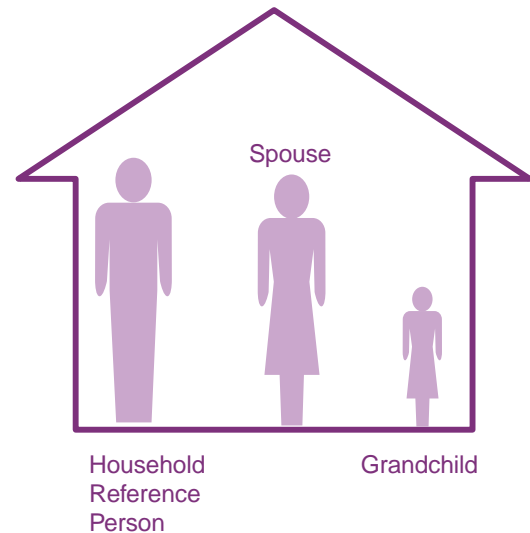
Describing Relationships to Reference Persons

The SIPP data files contain variables that identify household and family reference persons. They also contain variables that describe how each person in the sample is related to the household reference person.

Users should note that the identity of the household reference person can change from one month to the next; thus, the household description could also change.

Analysts can use other relationship variables on the files to identify a variety of family configurations, such as households containing three generations.

The *SIPP Users' Guide* discusses important differences in the 1996 and pre-1996 relationship variables.



Identifying Program Units

SIPP provides data for analyses involving program units for participants in transfer programs. SIPP records three characteristics regarding program participation:

- Whether the person is covered
- Who received the income or benefit
- The amount of the income or benefit

Coverage variables indicate whether a person is covered by a benefit directly or indirectly. For example, in a household receiving food stamps, the person who is the authorized recipient is identified as being covered directly.

Other members of the household are identified as being covered indirectly. Indirect recipients will have the same sample unit ID and current address ID as the primary recipient. **tip**

SIPP data also permit identification of members of common units within households, because most programs allow more than one program unit in a household. Members of common units can be identified by the sample unit ID and the authorized recipient variable. **tip**

Chapters 10–12 of the *SIPP Users' Guide* discuss specific variables related to program unit identification and exceptions to the rules for identifying program units.

SIPP *tip*

When a child receives a benefit, an adult will be the authorized recipient and will be flagged as not covered; the child will be flagged as covered. Except for WIC, no amounts of income or benefit are listed in the records of children under 15.

tip

Unlike most transfer programs, Medicare is a person-based program in which each participant is an authorized recipient. Thus, SIPP files do not carry additional authorized recipient variables on the files.

Identifying Movers and Household Composition Changes

When SIPP original sample members move, sometimes changes in household composition occur. The mover may acquire a spouse, a roommate, a child, or other new household members. It may be important for analysts to know about these household composition changes during a particular reference period.

To identify movers, analysts should look for changes in current address fields. Except in rare cases (e.g., merged households), movers' other basic ID variables—sample unit ID, entry address ID, and person number—remain the same. *tip*

Chapters 10–12 of the *SIPP Users' Guide* contain tables and explanatory text that illustrate how analysts can identify and track movers.

Identifying States and Metropolitan Areas

States. Even though it is possible to identify most states, SIPP was not designed to be representative at the state level. Therefore, SIPP data should not be used to produce state-level estimates.

Metropolitan Areas. Analysts can use variables in the core wave files to produce national estimates of the metropolitan population and to identify 93 Metropolitan Statistical Areas and Consolidated Metropolitan Statistical Areas.

Nonmetropolitan Areas. The Census Bureau recodes a small random sample of metropolitan households as non-metropolitan households to protect respondent confidentiality. Thus, SIPP data cannot be used to produce national estimates of the nonmetropolitan population.

SIPP *tip*

In the pre-1996 panels, when two SIPP households merged, or when one split but then recombined with new secondary sample members, some sample members may have received new ID variables. Because of the rarity of these cases, the 1996 Panel files do not include information about them.

Choosing Weights

SIPP samples different households and people at different rates. Consequently, analysts should use weights to reduce the likelihood of biased estimates of population characteristics.

SIPP data files include a number of alternative weights. The choice of the appropriate weight for an analysis depends on the population of interest—person, household, family, and so on.

Analysts need to ask:

1. Which sample or subsample of SIPP is the basis for the estimate?
2. What population does the sample represent?

To obtain weights, analysts should check the files they are using:

- Weights for each calendar month covered by a panel are in the core wave files.
- A single weight appears in the topical module files. **tip**
- Weights for calendar years are on the longitudinal files.

The source and accuracy statements that accompany the three types of files include suggestions about which weights to use and how to use them, as does Chapter 8 of the *SIPP Users' Guide*.

Income Topcoding

To protect the confidentiality of SIPP respondents, the Census Bureau topcodes very high incomes on the public use data files. New income topcoding procedures were instituted with the 1996 Panel.



SIPP *tip*

Before 1996, the weight on the topical module files is the person interview month weight for those who provided data for the module. In the 1996 Panel, the weight on the topical module file is the person cross-sectional weight for the fourth reference month.

1996 Panel

Unearned Income. When the total amount of asset income or of certain types of general income for a wave exceeds the established ceiling, the monthly amounts in excess of the monthly threshold are replaced by monthly topcode values. *tip*

Employment Income. Monthly employment income falls into three categories within SIPP:

- Wage and salary income
- Self-employed earnings
- Other worker arrangements

Each of these three sources was topcoded separately.

In the 1996 Panel, the method used to topcode employment income is based on the mean of reported unweighted amounts above the threshold in Wave 1 of the panel.

An algorithm was used to establish topcode values for 12 cells of different combinations of gender, race, and employment status. Each respondent's topcode value is assigned in accordance with his or her corresponding cell. *tip*

The topcode amounts established in Wave 1 of the 1996 Panel were used for all waves of the panel, with a wave adjustment, determined by formula, for inflation and real growth in earned income.

Pre-1996 Panels

In earlier panels, the topcode amount for the wave was \$33,332; thus, in most cases, the topcode amount for monthly income was \$8,333.

Income from various sources (multiple jobs, businesses, property) was not independently topcoded in the pre-1996 panels.

SIPP *tip*

Not all income sources are topcoded. For example, the amount of food stamp income is not topcoded. See Appendix B of the SIPP Users' Guide for a list of topcoded income variables in the 1996 Panel.

tip

Chapter 10 of the SIPP Users' Guide contains a discussion of the 1996 income topcoding method and examples illustrating its application.

Using Allocation Flags

As discussed earlier in the tutorial, the Census Bureau often imputes information when a person does not respond to the survey or to a particular question.

When a variable is imputed, the Census Bureau sets an allocation, or imputation, flag to identify the imputed variable. Variables selected for imputation vary across the three types of files.

Not all imputations are readily apparent, however.

Whole Record Imputation. Whole records were sometimes imputed with the Type Z procedure when person-level interviews were not successfully conducted. The variables needed to identify these records vary across the file types.

EPPFLAG and Little Type Z Imputation. In the 1996 Panel, the Census Bureau used special imputation procedures, known as EPPFLAG and little Type Z, for labor force items. The allocation flags for items imputed with these procedures will not indicate by themselves the imputation status of the items.

Analysts should read the discussion on allocation flags in Chapter 4 of the *SIPP Users' Guide* to learn how to identify items imputed with these special procedures.

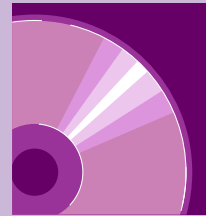
Composite Variables. Variables are imputed and the allocation (imputation) flags are set before the creation of composite variables, such as household and family aggregates. Since total household income is computed after person-level imputation has occurred, total household income may be based, in part, on imputed information. There will be no direct indication, though, on the records of other household members that any information on household income has been imputed.

Analysts should use the person-level imputation flags of all household and family members to identify aggregate amounts that include imputed values.

Using Core Wave Files

This section focuses on information specific to the core wave files.

- *File Structure*
- *Using the Data Dictionary*
 - 1996 Panel*
 - Pre-1996 Panels*
- *Identification/Description Variables*
 - Monthly Interview Status*
 - Identifying Persons*
 - Identifying Households*
 - Identifying Families*
 - Identifying Reference Persons*
 - Household Reference Person*
 - Family Reference Person*
 - Other Relationship Variables*
 - Program Units*
 - Movers & Household Composition Changes*
 - Identifying States & Metro Areas*
- *Family-Level Income Variables*
- *Topcoding*
- *Using Allocation Flags*
- *Weight Variables*



Structure of the Core Wave Files

In the first six SIPP panels, the core wave files were issued in person-record format. Beginning with the 1990 Panel, the core wave files have been issued in person-month format.

In the 1990–1996 Panels, one record per person exists for each month of the 4-month reference period that the person was in the sample. A person who was in the sample for all 4 months of the wave has four records.

If a person was not in the sample for the fourth month of the wave because he or she moved out of the country during the middle of the third month, for example, the file will contain three records. The third-month record for that person will contain information that was either imputed or collected by proxy from another household member.

The files also contain records for children under age 15 in sample households.

Using the Data Dictionary

The data dictionary is formatted to facilitate processing by user-written programs. The dictionaries in the 1996 Panel and earlier panels differ somewhat.

1996 Panel

- A “D” in the first column of the dictionary signifies that the line contains the variable name, size (i.e., the number of digits it contains), and the starting position.
- A “T” in the first column signifies that the line contains a short variable description that can be used by many software packages as a variable label.

```

SURVEY OF INCOME AND PROGRAM PARTICIPATION,
1996 PANEL WAVE 1 TOPICAL MODULE DATA DICTIONARY

DATA      SIZE  BEGIN
D SSUSEQ   5     1
T SU: Sequence Number of Sample Unit - Primary
  Sort Key
U All persons
V      1:50000 .Sequence Number

D SSUID    12     6
T SU: Sample Unit Identifier
  Sample Unit identifier This identifier is
  created by scrambling together the PSU,
  Segment, Serial, Serial Suffix of the
  original sample address. It may be used
  in matching sample units from different
  waves.
U All persons
V 000000000000:999999999999 .Scrambled Id

D SPANEL   4     18
T SU: Sample Code - Indicated Panel Year
U All persons
V

```

- A “U” in the first column signifies that the next words describe the universe.
- A “V” in the first column indicates that the next number and phrase describe one of the values of the variable.
- A blank in the first column denotes either a variable description or a comment.

Pre-1996 Panels

- A “D” in the first column of the dictionary signifies that the next few lines define the variable:
- The first line contains the variable name, size (i.e., the number of digits it contains), and the starting position.
- Succeeding lines contain a description of the variable.
- A “U” in the first column signifies that the next words describe the universe. *tip*
- A “V” in the first column indicates that the next number and phrase describe one of the values of the variable.
- An asterisk in the first column denotes a comment.
- A period (.) before a word denotes the start of the value label.

Identification/Description Variables

Monthly Interview Status

All core wave files issued in person-month format (1990 and subsequent panels) contain records only for persons whose respondent interview status was equal to 1. Thus, the monthly interview status variable can be safely ignored.

In the six earlier panels, core wave files were issued in person-record format. Users should check each person’s monthly interview status variables in these files.

SIPP *tip*

The universe definitions included in the data dictionaries before the 1996 Panel were not always accurate. Users of those panels should check the skip patterns in the actual survey questionnaires to determine which subset of respondents was asked each question.

Identifying Persons

To uniquely identify persons in the core wave files, analysts should use the following variables:

Variable Description	Pre-1996 Panels	1996 Panel
Sample unit ID	SUID	SSUID
Entry address ID	ENTRY	EENTAID (optional)
Person number ID	PNUM	EPPNUM

Chapter 10 of the *SIPP Users' Guide* provides illustrations of how to use these variables to identify individuals and learn when they first entered the SIPP sample.

Identifying Households

To uniquely identify households and group quarters in the core wave files, analysts should use the following two variables:

Variable Description	Pre-1996 Panels	1996 Panel
Sample unit ID	SUID	SSUID
Current address ID	ADDID	SHHADID

People with the same sample unit ID and current address ID live in the same household.

Identifying Families

By using several core wave variables and their associated numbering schemes, analysts can uniquely identify the following family configurations.

Primary Family (family containing the household reference person and all relatives living with him or her)

Variable Description	Pre-1996 Panels	1996 Panel
Sample unit ID	SUID	SSUID
Current address ID	ADDID	SHHADID
Family ID	FID	RFID

Primary Family Excluding Related Subfamilies (related subfamily: a family unit within the primary family whose members are related to, but do not include, the household reference person)

Variable Description	Pre-1996 Panels	1996 Panel
Sample unit ID	SUID	SSUID
Current address ID	ADDID	SHHADID
Family ID (excluding related subfamilies)	FID2	RFID2

Related Subfamilies Only

Variable Description	Pre-1996 Panels	1996 Panel
Sample unit ID	SUID	SSUID
Current address ID	ADDID	SHHADID
Family ID (for related subfamilies)	SID	RSID
Type of family	FTYPE	ESTYPE

Multigenerational Families

Variable Description	Pre-1996 Panels	1996 Panel
Sample unit ID	SUID	SSUID
Current address ID	ADDID	SHHADID
Family ID (excluding related subfamilies)	FID2	RFID2
Family ID (for both related and unrelated subfamilies)	SID	RSID

Identifying Household and Family Reference Persons

Analysts can use the following variables in the core wave files to identify the household reference person (the owner or renter of record) and family reference persons.

Variable Description	Pre-1996 Panels	1996 Panel
Household reference person	HREFPER	EHREFPER
Family reference person	FREFPER	EFREFPER

Describing Relationship to Household Reference Person

Analysts should note that there are two variables in the pre-1996 core wave files that describe how each person is related to the household reference person. One is an edited version of the other. The unedited version allows the analyst to describe more household relationships.

Variable Description	Pre-1996 Panels	1996 Panel
Relationship to household reference person	RRP RRPU (unedited)	ERRP

Chapter 10 of the *SIPP Users' Guide* contains tables that provide the values and value descriptions for these variables.

Describing Relationship to Family Reference Person

In the pre-1996 core wave files, analysts can use the variable FAMREL to identify the relationship of a person to the family reference person (such as spouse or child of family reference person).

The 1996 core wave files do not contain a variable that corresponds exactly to FAMREL. They do contain the variable ESFR (edited subfamily relationship), which is defined the same as FAMREL but applies only to related and unrelated subfamilies.

Identifying Other Relationship Variables

The core wave files contain many variables that describe household and family composition. [Link to a table from the SIPP Users' Guide that lists these variables.](#) Other material in Chapter 10 of the Guide provides more detail on these topics.

Note that in the following list of four of the relationship variables, just one parent is identified in files from panels before 1996.

Variable Description	Pre-1996 Panels	1996 Panel
Spouse	PNSP	EPNSPOUS
Parent	PNPT	
Father		EPNDAD
Mother		EPNMOM
Guardian	PNGDU	EPNGUARD

Identifying Program Units

Users will quickly note that the variable names for program units in the 1996 Panel are quite different from those in earlier panels.

[Link to a table from the SIPP Users' Guide that contains variable names for government transfer programs and health insurance programs in the core wave files.](#)

Questions about program units in the 1996 Panel were expanded in Waves 4 and 9 in response to replacement of the Aid to Families with Dependent Children (AFDC) program by a new program, Temporary Assistance for Needy Families (TANF). TANF provides a broader array of program types.

Identifying Movers and Household Composition Changes

Tables 10-14 and 10-15 in the *SIPP Users' Guide* provide examples of how to identify movers and changes in household composition in the core wave files.

In the rare cases of persons in merged households who were assigned new ID values, two records exist in the pre-1996 Panel core wave files for those persons when the move occurred after the first reference month. When the move occurred in the first reference month, only one record exists. Merged households cannot be identified in the 1996 Panel core wave files.

Identifying States and Metropolitan Areas

The purpose of including variables to identify states in the core wave files is to allow analysts to examine how state-level characteristics affect national estimates. As noted earlier, because SIPP data do not identify all states or uniquely identify nonmetropolitan residences, they should not be used to produce state-level or nonmetropolitan population estimates.

Variable Description	Pre-1996 Panels	1996 Panel
41 states, DC, and 3 combinations of 9 states	HSTATE	
45 states, DC, and 2 combinations of 5 states		TFIPSST
Metropolitan residences	HMETRO	METRO
93 MSAs and CMSAs	HMSA	TMSA

Family-Level Income Variables

Family-level income variables in the core wave files include the income of all related subfamily members. In other words, the Census Bureau treats primary family members, including related subfamily members, as one family when calculating family-level income amounts. The core wave files, however, also contain related subfamily income variables that aggregate the income of members of the same related subfamily.

Variable Description	Pre-1996 Panels	1996 Panel
Family income	FTOTINC	TFTOTINC
Related subfamily income	STOTINC	TSTOTINC

Analysts should keep these variable distinctions in mind when examining family income.

Topcoding

To protect respondents' confidentiality, the Census Bureau topcodes income and age-related variables in the public use files. See the information on topcoding income in the tutorial section SIPP Public Use Files.

Appendix B of the *SIPP Users' Guide* describes the Census Bureau's topcoding specifications for SIPP.


Using Allocation (Imputation) Flags

Almost all imputed person-level variables in the core wave files have allocation (imputation) flags.

In panels prior to 1996, the entire record was imputed if

- (1) MIS5 = 2 and MIS_j = 1 for j = 1, 2, 3, or 4 or
- (2) INTVW = 3 or 4.

The whole record was imputed in the 1996 Panel if EPPINTVW = 3 or 4.

EPPINTVW and INTVW describe the type of interview or noninterview that occurred with a person. 

Weight Variables

The core wave files include alternative reference month weights. Beginning with the 1996 Panel, SIPP files no longer include interview month weights.

Variable Description	1990–1993 Panels	1996 Panel
Reference month—final weight		
Person	FNLWGT	WPFINWGT
Household	HWGT	WHFNWGT
Family	FWGT	WFFINWGT
Related subfamily	SWGT	WSFINWGT
Interview month—final weight		
Person	P5WGT	
Household	H5WGT	

SIPP *tip*

Users should note that the codes for EPPINTVW and INTVW differ. Also, the method for identifying persons who were in the sample early in the wave but not at the time of the interview changed for the 1990–1993 Panels.

Using Topical Module Files

This section focuses on information specific to the topical module files.

- ***File Structure & Content***

- File Structure*

- General Content*

- Topical Module Vs. Core Wave Files*

- ***Variable Names, Reference Periods, & Sample Universe***

- Variable Names*

- Reference Periods & Sample Universe*

- ***Using the Data Dictionary***

- ***Identification/Description Variables***

- Monthly Interview Status*

- Identifying Persons & Households*

- Identifying Families*

- Household & Family Composition*

- Relationship to Household Reference Person*

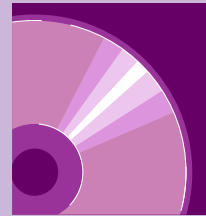
- Movers & Household Composition Changes*

- Identifying States & Metro Areas*

- ***Topcoding***

- ***Using Allocation Flags***

- ***Weight Variables***



File Structure and Content

Structure of the Topical Module Files

1996 Panel. The topical module files for the 1996 Panel contain one record for each person who was in the sample with a completed or imputed interview in the fourth month of the wave's reference period (the month before the interview).

Pre-1996 Panels. The topical module files for panels before 1996 contain one record for each sample member who was interviewed or for whom an interview was attempted during the interview month (month 5), not the fourth month of the reference period.

General Content of the Topical Module Files

Each topical module file contains data for all topical module subject areas administered during a given wave. The files also contain selected identification and demographic information from the SIPP core, making it possible to do some analysis of those files independently from core wave and full panel files.

If more detailed demographic information is necessary for an analysis, users can acquire that information by merging topical module files with core wave or full panel files, as discussed in the tutorial section Linking Files and in Chapter 13 of the *SIPP Users' Guide*.

Topical Module Vs. Core Wave Files

Topical module files differ from core wave files in key ways:

- The core wave files contain up to four records for each sample person in each wave (one record for each month of the wave the person was in the sample). The topical module files contain only one record for each SIPP sample member in each wave.
- For panels before 1996, topical module files include records for people whose entire households were not interviewed. Those people are excluded from the pre-1996 core wave files.

- As noted, the topical module files contain identification and demographic data also present in the core wave files. In the 1996 Panel, the values of those data correspond to month-4 data in the core wave file for the same wave.
- Prior to the 1996 Panel, however, the identification and demographic data in the topical module files correspond to data collected in the interview month (month 5), not to data in any of the 4 reference months. If any changes in those variables occurred between months 4 and 5, the values for the variables could differ between the core wave and topical module files.

Variable Names, Reference Periods, and Sample Universe

Variable Name

Prior to the 1996 Panel, some variable names were used in different topical module files for different variables, so the variable might change meaning depending on context.

Reference Periods and Sample Universe

Sample definitions and reference periods in topical modules vary across panels, across topical modules within panels, and even within topical modules. Analysts therefore need to pay close attention to those details in the topical module files they use.

1996 Panel. As noted above, most topical module questions were asked only of people in the sample during the fourth month of the wave's reference period. People who were members of SIPP households only at the time of the interview were not asked the topical module questions.

In addition, many questions applied only to the month of the reference period (month 4). However, some topical module questions—even entire topical modules—refer to longer periods of time.

Pre-1996 Panels. Most topical module questions were asked of people in the sample at the time of the interview (month 5). Thus questions were asked of some people who were not in the sample during the 4 previous months, the reference period for core questions in that wave. Consequently, to obtain core data that correspond to the topical module data, analysts must often merge core data from the subsequent wave.

Many topical module questions referred to “current” (interview month) conditions, although some asked about longer periods of time.

Using the Data Dictionary

The data dictionaries in the core wave and topical module files share the same format. The changes in format that occurred in the 1996 files apply to both core wave and topical module files. See the previous tutorial section, *Using Core Wave Files*.

Identification/Description Variables

Monthly Interview Status

Analysts should use data only for months in which the interview status variable has a value of 1.

1996 Panel. Only one interview status variable appears in the 1996 topical module files (EPPMIS4). Because those files contain records only for people who were in the sample, EPPMIS4 always equals 1 and can be safely ignored.

Pre-1996 Panels. The topical module files for these panels are different. There are five interview status variables (PP-MISx), one for each of the reference months and one for the interview month itself (PP-MIS5). Questions were asked only of sample members whose interview status in the interview month was equal to 1.

```

SURVEY OF INCOME AND PROGRAM PARTICIPATION,
1996 PANEL WAVE 1 TOPICAL MODULE DATA DICTIONARY

DATA      SIZE  BEGIN
D  SSUSEQ   5    1
T  SU: Sequence Number of Sample Unit - Primary
   Sort Key
U  All persons
V    1:50000 .Sequence Number

D  SSUID    12    6
T  SU: Sample Unit Identifier
   Sample Unit identifier This identifier is
   created by scrambling together the PSU,
   Segment, Serial, Serial Suffix of the
   original sample address. It may be used
   in matching sample units from different
   waves.
U  All persons
V  000000000000:999999999999 .Scrambled Id

D  SPANEL   4    18
T  SU: Sample Code - Indicated Panel Year
U  All persons
V

```

Identifying Persons and Households

The variables analysts should use to track persons and households are the same in both the core wave and topical module files except that the variable name of the sample unit ID in the pre-1996 topical module files is ID (see previous tutorial section, *Using Core Wave Files*).

Identifying Families

The variables analysts should use to track families in the topical module files are also the same as those used in the core wave files, except that the topical module files for the 1996 Panel do not contain the variable needed to determine whether all subfamily members are members of the same subfamily. To determine that, an analyst must merge the RSID variable from the month-4 records in the core wave file.

Describing Household and Family Composition

The topical module files contain fewer variables describing household and family composition than do the core wave files. [Link to a table with the topical module variables.](#)

Analysts wanting more details can merge additional variables from the core wave or full panel files.

Describing Relationship to Household Reference Person

The 1996 Panel core wave and topical module files contain the ERRP variable, which analysts can use to describe relationships to the household reference person. The pre-1996 topical module files contain only RRP, the edited version of the variable used to describe relationships to the household reference person. When a fuller description is needed, analysts can merge the unedited variable (RRPU) from the core wave files.

Identifying Movers and Household Composition Changes

The procedures for identifying movers and household changes are the same in the topical module files and the core wave files. Chapter 11 of the *SIPP Users' Guide* describes and illustrates the procedures in text and tables.

In the rare cases of merged households where persons may have two sets of ID values, the pre-1996 topical module files contain records for those persons only after the move. Analysts must use the core wave file records to identify those persons before the move. Persons in merged households cannot be identified in the 1996 Panel files.

Identifying States and Metropolitan Areas

The same caveat that applies to the core wave files also applies to the topical module files regarding state identification: SIPP was not designed to be representative at the state level and should not be used to produce state-level estimates.

The following variables for identifying states were included in the topical module files only to allow analysts to examine how state-level characteristics affect national estimates.

Variable Description	Pre-1996 Panels	1996 Panel
41 states, DC, and 3 combinations of 9 states	State	
45 states, DC, and 2 combinations of 5 states		TFIPSST

The topical module files do not contain any variables identifying metropolitan areas. Analysts needing that information must merge it from core wave files.

Topcoding

The topcoding procedures used in the topical module files are similar to those used in the core wave files. In general, topcodes for continuous variables, such as income, that apply to the total population include at least $\frac{1}{2}$ of 1 percent of all cases. For income variables that apply to subpopulations, topcodes include either 3 percent of the appropriate cases or $\frac{1}{2}$ of 1 percent of all cases, whichever is higher.

Characteristics frequently topcoded in the topical module files include income and expense values, including those for a broad range of assets and liabilities. The documentation for these variables indicates whether the values are topcoded and the value ranges for the variables.

Using Allocation (Imputation) Flags

As in the core wave files, there is an allocation (imputation) flag for almost all of the person-level variables that are imputed.

There are two ways to identify cases with edited or imputed data in panels prior to 1996: The entire record was imputed if:

- (1) PP-MIS5 = 2 and PP-MISj = 1 for j = 1, 2, 3, or 4 or
- (2) INTVW = 3 or 4.

The whole record was imputed in the 1996 Panel if EPPINTVW = 3 or 4.

Weight Variables

The topical module files contain one weight variable:

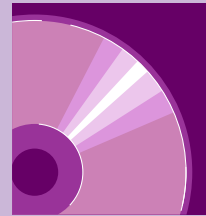
- WPFINWGT in the 1996 Panel—the person cross-sectional weight for the fourth reference month
- FINALWGT in the pre-1996 Panels—the person interview month weight for people who provided data for a topical module

Using the 1990–1993 Full Panel Files

This section focuses on information specific to the full panel files.

Because the 1996 full panel file is not yet available, the information in this section applies only to the 1990–1993 full panel files.

- *File Structure*
- *Using the Data Dictionary*
- *Aligning Data by Month*
- *Identification/Description Variables*
 - Monthly Interview Status*
 - Identifying Persons*
 - Identifying Households*
 - Identifying Families*
 - Family & Household Composition*
 - Identifying Program Units*
 - Movers & Household Composition Changes*
 - Identifying States & Metro Areas*
- *Income Variables*
 - Family-Level Income*
 - Unearned Income*
 - Topcoding*
- *Using Allocation Flags*
- *Weight Variables*



Structure of the 1990–1993 Full Panel Files

The full panel files contain one record for each person who was ever in the SIPP sample for that panel. This is true even if the person was in the sample for just 1 month. Full panel files contain records for children and for people who entered the sample after the first wave.

Within each record, variables correspond to the information collected in the core interviews. However, some core items, including some constructed variables, are not included on the full panel files. No items from the topical module files are on the full panel files. *tip*

Using the Data Dictionary

The format of the data dictionary for the 1990–1993 full panel files is similar to that used for the pre-1996 core wave and topical module files except that two extra fields are added to lines with a “D” in the first column. These two fields denote:

- The number of occurrences of the variable (for example, some questions were asked each wave of the panel, and some questions were asked each month of the panel)
- The number of digits for each occurrence of the variable *tip*

Aligning Data by Calendar Month

Analysts often find it useful to realign SIPP data by calendar month rather than reference month. For example, to analyze data for a specific calendar year or fiscal year, SIPP users must realign the data.

There are various approaches for realignment. In each case, the analyst must use the technical documentation to determine the reference period for each rotation group of the panel. Analysts also need to apply the mapping from reference month to calendar month for each person included in the analysis.

SIPP *tip*

Analysts familiar with the core wave files should be careful when using the full panel files. Important information about families, unearned income, and other key topics is coded and/or organized differently in the two file types.

tip


The data dictionary for the 1992 full panel file has a line labeled with an “R” in column 1. This line provides value ranges for the variable. Also, fields in lines beginning with a “D” vary somewhat from “D” fields in other full panel files.

Chapter 12 of the *SIPP Users' Guide* contains an algorithm that realigns data by calendar month. In the algorithm, the first step realigns the months; the second step initializes each monthly variable to distinguish the months in which the variable is not relevant. Finally, the algorithm realigns the input data to be based on the calendar month.

[Link to the algorithm.](#)

Identification/Description Variables

Monthly Interview Status

In the full panel files, the monthly interview status variable (PP-MIS), which helps determine whether data for a person in a given month should be used, occurs once for each reference month of the panel. Analysts should use data only for months in which the interview status variable has a value of 1. 

Identifying Persons

To uniquely identify a person in the 1990–1993 full panel files, analysts should use the following three variables:

Variable Name	Description
PP-ID	Sample unit ID
PP-ENTRY	Entry address ID
PP-PNUM	Person number

PP-ID is a random recode of three variables in the Census Bureau's internal files. The variables are omitted from the public use files to protect the confidentiality of respondents.

Identifying Households

To uniquely identify households and group quarters in the 1990–1993 full panel files, analysts should use the following variables:


Variable Name	Description
PP-ID	Sample unit ID
HH-ADDID _{<i>i</i>}	Current address ID in the <i>i</i> th month
PP-MIS _{<i>i</i>}	Person's interview status in the <i>i</i> th month

SIPP *tip*

Analysts should be careful not to confuse the monthly interview status variable with the interview status variable (PP-INTVW).

Because household composition changes from one month to the next, it is generally not possible to construct “longitudinal households.” For a given person, analysts should evaluate the characteristics of the household each month. Characteristics should cover only those people who reside together in each specific month.

Identifying Families

Unlike the core wave files for the 1990–1993 Panels, the corresponding full panel files do not contain family identification variables (e.g., FID, FID2, and SID). Analysts needing family identification variables must either merge them from the core wave files or create them. Because family composition can change over time, these are monthly variables. 

[Link to an algorithm that provides one approach to creating functional equivalents of the variables on the core wave files.](#)

Describing Family and Household Composition

Analysts can use the household ID variables and the variables created by the “family” algorithm to group people into the same family and subfamily groups that appear in the core wave files. However, the actual values assigned by this algorithm to these variables generally will not equal the values found in the variables from the core wave files.

The 1990–1993 full panel files also include nine additional variables that can be used to identify relationships to reference persons and a variety of household configurations, including households containing three generations.

[Link to a table containing the nine household description variables.](#)

Identifying Program Units

The 1990–1993 full panel file information on participation in health insurance and government transfer programs differs in some ways from the corresponding core wave file information.

SIPP *tip*

Beginning with the 1991 Panel, a new missing wave imputation procedure was applied to the full panel files: data were imputed for people with missing data for a wave but with valid data for the two adjacent waves. For these people, merging the core wave family ID variables is not an option.

1. In the full panel files, the authorized recipient variables do not use the entry address and person number values. Instead, they use the sequence number of the person within the sample unit (PP-RCSEQ) to identify authorized recipients. For example, the authorized food stamp recipient is the person for whom FS-PIDXi in month *i* equals PP-RCSEQ.
2. The variables used to identify members of a common program unit in a given month (*i*) can be identified with the following three variables:
 - Sample unit ID—PP-ID
 - Person's interview status in month *i*—PP-MIS_{*i*}
 - Authorized recipient variable in month *i*
3. Unlike the core wave files, the full panel files have no coverage variable indicating whether the child, adult, or both were covered by SSI. If needed, that information can be acquired from merges with the core wave files.

Identifying Movers and Household Composition Changes

The procedures for identifying movers and household changes are essentially the same in the 1990–1993 full panel files as in the corresponding core wave and topical module files. In the rare cases of persons in merged households who were assigned new ID values, the full panel files contain two full panel records for those persons.

Chapter 12 of the *SIPP Users' Guide* describes the procedures for tracking movers in the 1990–1993 full panel files.

Identifying States and Metropolitan Areas


States. SIPP is not designed to allow analysts to produce state-level estimates. A state variable (GEO-STE) is included on the 1990–1993 full panel files to allow examination of how state-level estimates affect national-level estimates. GEO-STE identifies 41 individual states and the District of Columbia; the remaining 9 states are suppressed into three groups.

A user could apply the state-specific eligibility criteria for a means-tested program to arrive at a national estimate of the number of people eligible for the program.

Metropolitan Areas. The full panel files do not contain any variables identifying metropolitan areas. Analysts needing that information must merge it from the core wave files.

Income Variables

Family-Level Income Variables

The family-level income variables in the full panel files, like those in the core wave files, include the income of all related subfamily members. However, unlike the core wave files, the full panel files do not contain any subfamily income variables. If family income variables are needed that do not pool related subfamilies with primary families, those income variables must be created. 

Unearned Income Variables

Analysts need to be aware that the Census Bureau organizes the unearned income variables differently in the core wave and full panel files.

In the full panel files, 10 variables on each person's record identify up to 10 different sources of unearned income. For each source identified, there is a corresponding amount variable.

When using the unearned income fields in the full panel files, analysts often find it helpful to realign the unearned income into new income-specific variables.

[Link to an algorithm that demonstrates how to create income-specific variables.](#)


Income Topcoding

Income topcoding procedures in the 1990–1993 full panel files are the same as those used in the core wave files of the 1990–1993 Panels.

SIPP *tip*

Unpooled income variables can be created by looping over persons with PP-MIS_j of 1 and with common PP-ID, HH-ADDID_j, FID2, and SID_j for each month.

Using Allocation (Imputation) Flags

The edit and imputation procedures used for the 1990–1993 full panel files differ from those used for the corresponding core wave files. The procedures for the full panel files make use of a full set of longitudinal data for a person, in contrast to a maximum of 4 months of observations that can be applied to the core wave files. The procedures applied to the core wave files make greater use of cross-observation imputation methods than do those applied to the full panel files. 

Two sources identify whether information has been imputed in the 1990–1993 full panel files:

1. Beginning with the 1991 Panel, all data for a wave are imputed if a person was not successfully interviewed in one wave but had complete information (from either a successful interview or a proxy interview) in the two adjacent waves. In those cases, the value of WAVFLG will be greater than zero and INTVW will be 3 or 4.
2. Imputation flags appear for a limited set of variables, including earned income, asset income, and unearned (transfer) income variables.

Weight Variables

The 1990–1993 full panel files include:

- The calendar year weights—FNLWGTs
- The full panel weight—PNLWGT

The number of calendar year weights corresponds to the duration of the panel.

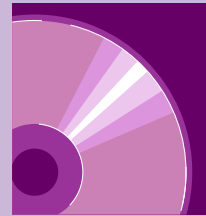
SIPP *tip*

The edit and imputation procedures applied to the core wave files from the 1996 Panel make greater use of prior wave information than procedures used in earlier panels.

Linking Core Wave, Topical Module, and Full Panel Files

This section describes reasons and procedures for linking files, including suggestions for handling nonmatches.

- *Reasons for Linking Files*
- *Procedures for Linking Files*
 - Three Basic Steps*
 - Six Types of Merges*
- *Descriptions of the Six Types of Merges*
 - Within a Core Wave File*
 - Two or More Core Wave Files*
 - Core Wave and Full Panel Files*
 - Two or More Topical Module Files*
 - Topical Module and Core Wave Files*
 - Topical Module and Full Panel Files*
- *Nonmatches and Other Anomalies*
 - Entering and Exiting the Population*
 - Sample Attrition*
 - Missing Wave Imputation*
 - Merged Households*



Reasons for Linking Files

Often, a single SIPP data file will not contain all the information needed for a project. In those cases, analysts may need to merge data from another file or link two or more files. For example, analysts often link SIPP files for the following reasons:

- Data for a single calendar reference month are often contained on two different core wave files.
- In the pre-1996 Panel files, data covering a single calendar year are often on files from two or even three different panels.
- Analysts may need to merge topical module data with core wave data.
- Analysts may need to link core wave files for a longitudinal analysis if the full panel file has not been released or if the variables of interest are not available in the longitudinal file (for pre-1996 files).

Procedures for Linking Files

In this tutorial section, and in Chapter 13 of the *SIPP Users' Guide*, procedures for linking person records across files are described. Procedures for linking households or families are problematic when working with longitudinal data—because unit composition changes over time—and are therefore not discussed.

Three Basic Steps

To link files, analysts need to:

1. Create data extracts from each file to be linked.
2. Sort the files in common order by using identified variables as match keys.
3. Merge the files.



Depending on the planned analysis and software used, analysts choose to create final files either in person-month format, reflecting the 1990 and later core wave files, or in person-record format.

Six Types of Merges

SIPP users commonly merge files in the following ways:

1. Within a core wave file
2. Two or more core wave files
3. Core wave and full panel files
4. Two or more topical module files
5. Topical module and core wave files
6. Topical module and full panel files

Information about the ID variables needed for the six types of merges is provided in Chapter 13 of the *SIPP Users' Guide*.

Descriptions of the Six Types of Merges

Merges Within a Core Wave File

Core wave files have one record per person per month. Linking within a core wave file transforms the files into a single wide record per person—the format used for core wave files before the 1990 Panel.

Chapter 13 of the *SIPP Users' Guide* describes two approaches for this linking process. Programmers using third-generation languages such as FORTRAN and PL/1 use one approach. Programmers using fourth-generation languages such as SAS and SPSS typically use the second approach. **tip**

Merging Two or More Core Wave Files

There are two reasons to link two or more core wave files:

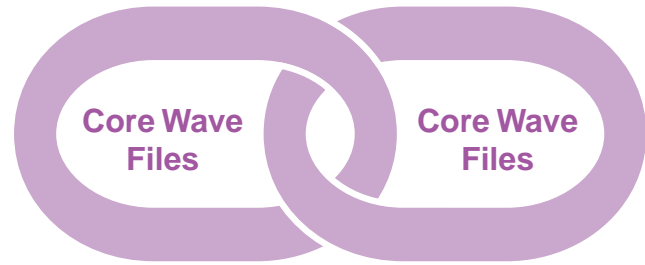
1. To create an analysis file with more than 4 months of information for each person

SIPP *tip*

Chapter 13 of the SIPP Users' Guide contains sample SAS code for changing core wave files from person-month format to person-record format.

2. As a step in merging core wave data with data from another file type

To create a final-analysis file in person-month format from two or more waves, the analyst simply needs to sort and interleave the files. Refer to Chapter 13 of the *SIPP Users' Guide* for the necessary variables that will ensure a proper sort. To create files in person-record format with just one record per person, analysts first need to interleave files to create the person-month-format file. Analysts can then apply procedures for merging within a core wave file.



Effects of Editing and Imputation. Analysts should be careful when creating their own longitudinal databases from core wave files in the pre-1996 panels. All edits and imputations in a wave were independent of those used in other waves; thus, data across waves may be inconsistent. For basic demographic information, it is generally safe to assume that the most recent data are correct. *tip*

Weights. Analysts should note that the sample weights included on the core wave files are calendar month specific. These weights may not be appropriate for longitudinal analyses with linked core wave files.

Merging Core Wave and Full Panel Files

This procedure is not used very often because the two files contain the same information for the most part. However, some core information appears only on the core wave files, making it necessary at times to merge the core wave and full panel files.

To link data from the two file types, analysts should do the following:

1. Create data extracts from the core wave and full panel files.

SIPP *tip*

New edit and imputation procedures that make use of prior wave data were used in the 1996 Panel to improve data consistency. Logical inconsistencies will still exist in the 1996 Panel files among reported items that were not longitudinally edited (basic demographic characteristics were longitudinally edited).

2. Put the extracts into the same format.
3. Sort the extracts in the same order.
4. Merge the extracts, creating the final file. *tip*

Chapter 13 of the *SIPP Users' Guide* discusses specific steps involved in transforming the data. It also includes sample SAS code.

Analysts should note that edit and imputation procedures differ for some variables. In addition, starting with the 1991 Panel, SIPP missing wave imputation procedures have created a situation in which data may be present in the full panel files but not in the core wave files.

Merging Two or More Topical Module Files

Analysts may wish to study the relationship between subject areas covered by different topical modules. For example, they might want to study the relationship between education and training history as reported in the second wave of the 1996 Panel and employment history as reported in the first wave of the 1996 Panel. In that case, they will need to link topical module files. In some panels, all of those data are reported in the same wave and no merge is necessary.

Topical module files are relatively simple to merge because they all have the same format (one record per person). Also, the ID variables are the same across files, except that the names for those variables differ between the 1996 and pre-1996 files (e.g., SSUID vs. ID). Nevertheless, analysts need to be cautious:

- Prior to the 1996 Panel, a variable name sometimes was used in different topical module files for different variables.

SIPP *tip*

Key variables have different names in the core wave and full panel files. Analysts should check the technical documentation to make sure that they are matching information as they intend.



- Not all people with records in one topical module file will have records in another topical module file. Household composition may have changed from one wave to the next, and this will be reflected in the topical module files. In addition, nonmatches might occur because of nonresponse. Also, universes for topical modules may differ.
- The substantial number of nonmatches across topical modules complicates the choice of weights. Analysts might instead want to use one of the weights from the full panel files.

Analysts wishing to measure change with data from the topical module files should be careful because of changes in measurement over time.

In addition, apparent changes across pre-1996 topical modules could be due to real changes reported by the respondent or to edit and longitudinal inconsistencies.

Merging Topical Module and Core Wave Files

It is sometimes necessary to merge topical module files with data from the core wave files. Analysts should be careful when selecting which core wave file to use—some topical modules sought information about the interview month, while the core wave files contain information about a different reference month.



Topical module files have one record per person, while core wave files have as many as four. Therefore, three options exist for merging topical module and core wave files:

1. Select a single month from the core wave files.
2. Spread the topical module data across all records from the core wave file, which results in a final file in person-month format.

3. Create a single record for each person from the core wave file and merge the topical module data to that record.

Analysts should execute the following steps:

1. Create an extract from the core wave file.
2. Apply the appropriate algorithm, as shown in Chapter 13 of the *SIPP Users' Guide*.
3. Sort the core wave extract by using the sort keys that uniquely identify people in the core wave file.
4. Create an extract from the topical module, and sort.
5. Merge the core wave extract with the topical module extract and sort. Sort keys will be different for the 1996 Panel and previous panels. *tip*

Merging Topical Module and Full Panel Files

This procedure applies to panels prior to 1996. There are times when analysts will want to merge topical module and full panel files. For example, if the full panel weights are needed for the planned analysis, they must come from the full panel files. *tip*

The full panel files contain a record for every person who was ever a member of a SIPP household. Therefore, every person with a record in a topical module file should have a record in the full panel file. Analysts working with a person-month file may nonetheless find nonmatches.

For this type of linkage, analysts should carry out the following steps:

1. Create an extract from the full panel file.
2. If the person-month format is desired, apply the appropriate algorithm (see Chapter 13 of the *SIPP Users' Guide*), but rename the ID variables to match those used in the topical module files.
3. Sort the full panel extract.

SIPP *tip*

In the pre-1996 Panels, there will likely be nonmatches between the file types because people who were present in the interview month (topical module files) may not have been present during any of the previous 4 months (core wave files).

tip

The edit and imputation procedures used with the full panel files are believed to introduce less error than the procedures used with the core wave files. Thus, when the same core items are available from the core wave and full panel files, analysts may prefer to use the full panel files.

4. Create an extract from the desired topical module file, and sort.
5. Merge the two extracts by using the appropriate ID variables.

Nonmatches and Other Anomalies

SIPP follows a group of people over a period of time. Original sample members are followed throughout the time period unless they die or leave the sample universe by moving to an ineligible location, such as a nursing home, a military barracks, or another country. Secondary sample members are part of SIPP only when they live with an original sample member.

Nonmatches occur when analysts merge across waves for any file types. Respondents may be in one data file and not another for a number of reasons:

- Original sample members move to (or back from) ineligible locations or drop out of the sample but not the sample universe.
- Secondary sample members move into or out of the sample.
- The person is a newborn.
- Missing wave data imputed in the full panel file is not in the core wave or topical module files.
- The person was in a merged household and received new ID information.

Entering and Exiting the Population

There is a fundamental distinction between situations in which people leave the sample because they leave the SIPP sample universe and situations in which they leave the sample but are still part of the population.

In general, when nonmatches occur because of people entering or exiting the population of the sample, data should not be imputed and weights should not be adjusted for the period of their absence.

Analysts can employ a number of strategies to deal with these nonmatches:

- They can drop leavers from the sample entirely and not adjust the weights of the retained cases. The remaining sample now represents the population that existed at both Time 1 and Time 2. *tip*
- Event-history models can also be used, with a person's exit from the population as one of the competing outcomes.

Sample Attrition

Sample attrition occurs when people leave the sample but remain part of the population represented by the sample. Several options exist for handling such cases. Analysts can choose to:

- Impute the missing data
- Eliminate cases with missing data and poststratify the weights for the retained cases
- Use a subset of cases with complete data and Census Bureau–provided weights
- Use other missing data methods to provide estimates and standard errors *tip*

Missing Wave Imputation

Beginning with the 1991 Panel, the Census Bureau has applied a missing wave imputation procedure to full panel files. Persons with missing data for one wave but complete data for two adjacent waves have data imputed.

If these cases were person-level nonrespondents who had data imputed with different methods in the core wave files, the data in their full panel and core wave records will differ. Other persons may have data for the missing wave only in the full panel file. For a complete explanation of the handling of missing wave data in SIPP, refer to the study “Compensating for Missing Wave Data in the Survey of

SIPP *tip*

Dropping leavers from the sample is simple to do, but analysts then cannot draw inferences about the part of the population that has left. For example, the economic profiles of people leaving the sample to enter prison or a nursing home will likely differ from the profiles of those who remain in the sample.

tip

All of the methods for handling sample attrition require caution. Chapter 13 of the SIPP Users' Guide presents an in-depth discussion of the possible pitfalls.

Income and Program Participation” by Williams and Bailey, which can be accessed from the SIPP home page under Publications.

The correct procedure for dealing with these nonmatches depends on which weights will be used.

- If weights come from the core wave or topical module files, analysts should drop observations from the full panel files that are not present in the cross-sectional files.
- If weights come from the full panel file, the Census Bureau suggests using the procedures for sample attrition.

Merged Households

Nonmatches can occur when the Census Bureau changes ID numbers for sample members. In panels before 1996, there were two very rare occasions when this happened. The first was when two separate sampling units with original sample members merged together, perhaps because of a marriage. The Census Bureau changed the identification information of one set of original sample members to agree with the other set.

The second instance occurred when a SIPP household split into new households, gained new secondary sample members in each, and later recombined with the secondary sample members coming along. In the recombined household, the secondary sample members from one of the earlier split households were assigned new person numbers.

Different file types recorded this information differently. Chapter 13 of the *SIPP Users' Guide* discusses this situation in-depth and tells how analysts can search the core wave file for these people. Analysts can then change the identification information, duplicate and merge the records, or treat the person with the new identity as two people, as is done in the full panel files.

Analysis Example

The following questions and answers illustrate typical SIPP analysis tasks—for example, choosing panels and interview months, understanding file structure and definitions of terms, recoding/creating variables, and merging files.

NOTE: [BLUE](#) INDICATES A HYPERLINK TO THE CORRESPONDING VARIABLES AT THE END OF THE DOCUMENT.

QUESTION

I want to study adult female labor force participants with young children (5 years old or younger) in the family and determine whether they ever participated in the Food Stamp program. I would like to use the 1986, 1991, and 1996 SIPP Panels to compare that population at 5-year intervals. How would I do that?

ANSWER

PART 1: Within the SIPP panels, which interview should I choose?

To answer the part of the question concerning past food stamp reciprocity, the analyst needs to use the Reciprocity History module. In SIPP 1986 and 1991, this module occurred in the second interview. In SIPP 1996, this module was asked in the first interview. To simplify this example, take the core information from the same interview as the Reciprocity History module. If it is desirable to use a different interview, the analyst needs to add up [food stamp coverage](#) flags across the intervening interviews.

Depending on the year of the SIPP panel, the data will look different. In the 1984–1988 Panels, the topical module and core files are combined on one data set and are in person-record format (each person has one record). In the 1990–1993 Panels and the 1996 Panel, the core and topical module information is separated and the core file is in person-month format (a record exists for each reference month of each interview). In general, the topical module refers to the [last month of the interview](#) ([reference month](#) 4).

The specific panels that were chosen are typical of different panel years of SIPP. The 1984–1988 Panels can be viewed as one group; the 1990–1993 Panels as a second group; and the 1996 Panel as a third, separate group.

PART 2: How can I study adult females in the labor force?

To study adult females, the analyst needs to confirm that each person was [interviewed](#), the [age](#) corresponds to an adult, and the [sex](#) is female. When everyone in the sample meets these three criteria, the sample will include only adult females.

Labor force participation is defined to include persons either [working](#) or [looking](#) for a job. If a person is not working and not looking, the person is a nonparticipant in the labor force. SIPP allows the analyst to look at all aspects. Because SIPP interviews cover 4 months of information, the analyst could choose any month or all months in defining labor force status. In this example, a person is in the labor force if she worked or looked for work during the [last month of the interview](#). A variable should be created indicating [labor force status](#).

Households in SIPP are interviewed every 4 months. However, each household is not interviewed at the same time. The households are divided into four groups ([rotation groups](#)), and one group is interviewed in a given month. This rotation procedure is used because the total number of interviews to be conducted for the 4-month period is too large to do at one time.

If the analyst uses the [last month in the interview](#), the data will represent an average over 4 months. A researcher could also use a specific calendar month, instead of the average over 4 months. Chapter 12 of the *SIPP Users' Guide* discusses the general approach for determining [calendar months](#). However, if calendar months are used, the time frame may not correspond to the typical time frame for the topical modules.

PART 3: How do I determine that a person has young children in the family?

To answer this part of the question, the analyst needs to create an ID variable that captures how many young children are in the family. The concept of “family” needs to be addressed because the Census Bureau allows multiple options. The default option defines a family as consisting of all household persons related by blood, marriage, or adoption. This definition allows for multigenerational units within the family.

The alternative option allows the analyst to split the larger family groups into smaller ones. These multi-unit families can be identified for the primary family only, that is, the family group that contains the household reference person. In the multi-unit family, the group of persons immediately related to the reference person (such as spouse or unmarried child) can be separated from other relatives, provided the

other relatives have relatives present. The classic example is a two-parent family (one of whom is the household reference person) with an adult female child who has a child or spouse of her own living in the household. The default definition treats this entire group as one family, and the group is referred to as a primary family. The alternative definition generates two families, one consisting of the husband and wife and the other consisting of the adult child and her child or spouse. The latter family is referred to as a related subfamily.

In general terms, a related subfamily is a family unit within the primary family whose members are related to, but do not include, the household reference person. As noted earlier, examples include a married daughter or son and spouse (with or without children) or a single parent with a child related to and living in the home of the household reference person.

Households may also include unrelated subfamilies—families living in the household whose members are not related to the household reference person.

Because people can enter or leave the household, the persons constituting a family can change each month.

For determining [poverty](#), the Census Bureau uses the inclusive definition of family. This choice was based on the concept of family dependency. The “[inclusive family ID](#)” should be used to get the poverty status for the family and any other family recodes that might be desired. The question presented concerns labor force status of females with young children in the family. This makes the Census Bureau’s definition more appropriate.

If the question focused only on the labor force status of *mothers* who have young children, and not all female adults who have young children, the above approach would not be appropriate. Instead, the analyst would need to create a “[modified family ID](#)” variable. This ID would take into account the need for related subfamilies to have a separate family ID that is different from the primary family ID.

Another approach to identifying the children would be to use the variables that point to the [parent](#) or [guardian](#) and create new identifying variables. The variables an analyst would use with this more-complicated approach are discussed in Chapter 10 of the *SIPP Users’ Guide*.

Because the topical modules typically refer to the last month in an interview, this example fixes the family structure at the [last month in the interview](#).

Once the appropriate family ID variables are created, a [counting](#) program can be used to add up the number of young children associated with each family ID variable.

PART 4: How do I get the information on whether or not these women have participated in the Food Stamp program?

The topical modules contain information that is not asked at every interview. The Reciprocity History module contains information on [past participation in food stamps](#). Information on current food stamp participation is contained in the core data, and participation is identified with [food stamp coverage](#) flags. These flagged variables are in the core data file and should be kept with any other variables of interest (demographics, population weight, etc.) discussed in Part 2. If there are data indicating past or present participation, then the person has participated in the Food Stamp program.

If researchers need to focus on an interview period that occurred after the Reciprocity History module, they would have to gather the information from subsequent interviews.

PART 5: How do I combine the information from Part 2, Part 3, and Part 4?

To combine the data from the various parts, the analyst needs to create various identifiers. Part 2 and Part 4 can be combined by making [person ID](#) variables. These variables will be used to merge the two datasets. In Part 3, the counting program produced two variables: "[inclusive family ID](#)" and the number of young children in the family. On the new data set, the "[inclusive family ID](#)" needs to be created so that it can be merged with Part 3.

In the resulting data set, keep only the observations that have information from each part and that have a positive number of young children. If labor force status or number of young children or ever-received food stamps is blank, then delete the observation.

This will leave a final data set of females who have young children in the family. If the analyst wants to focus on females participating in the labor force, the analyst would use the [labor force status](#) variable to select only participants. In addition, each observation has information on whether the female has ever participated in the Food Stamp program.

For SIPP 1986 variables, the “4” in a variable name represents the variable in [the last month in the interview](#).

[Age](#)

1986: AGE_4

1991: AGE

1996: EAGE

[calendar month](#)

Within an interview, there will be only 1 calendar month that the different rotation groups have in common. If an analyst wants to use a different month from the one that is common, then different interviews would have to be combined. Further modifications to this example would need to be made. For instance, the last month of the interview would not determine the sample. Instead, it would be the common calendar month. The changes below are the changes necessary for doing a calendar month estimate within the wave containing the Reciprocity History. These adjustments give: May of 1986, May of 1991, and March of 1996.

1986: If the [rotation group](#) equals 2 then use the “4” variables.

If the [rotation group](#) equals 3 then use the “3” variables.

If the [rotation group](#) equals 4 then use the “2” variables.

If the [rotation group](#) equals 1 then use the “1” variables.

1991: Keep the Person/Month record that meets the conditions below.

If the [rotation group](#) equals 2 and the [reference month](#) equals 4.

If the [rotation group](#) equals 3 and the [reference month](#) equals 3.

If the [rotation group](#) equals 4 and the [reference month](#) equals 2.

If the [rotation group](#) equals 1 and the [reference month](#) equals 1.

1996: Keep the Person/Month record that meets the conditions below.

If the [rotation group](#) equals 1 and the [reference month](#) equals 4.

If the [rotation group](#) equals 2 and the [reference month](#) equals 3.

If the [rotation group](#) equals 3 and the [reference month](#) equals 2.

If the [rotation group](#) equals 4 and the [reference month](#) equals 1.

[counting](#)

In this context, a counting program counts the children 5 years old or younger. Initially, set the counter to zero. Within a family, count each person that is in the age group (count=count+1 when age less than 6). Keep only the last family record because that record will contain the total number of young children. At the end of this program, there will be one record per family. Each record will contain the family ID and the family recode for young children.

[inclusive family ID](#)

These variables make a unique Census-style family ID. Unrelated subfamilies receive a family sequence number that is distinct from the householder's family.

1986:	SS_ID	H4_ADDID	F4_NUMBR
1991:	SS_ID	ADDID	FID
1996:	SSUID	SHHADID	RFID

[interviewed](#)

When the population weight is greater than zero, the interview is considered "good."

1986:	FNLWGT4
1991:	FNLWGT
1996:	WPFINWGT

[food stamp coverage](#)

If the variable equals one, then the person is covered by food stamps.

1986:	FOODSTP4
1991:	FOODSTP
1996:	RCUTYP27

[guardian](#)

This is the person number of the guardian.

1986:	PNGDU
1991:	PNGDU
1996:	EPNGUARD

last month in the interview

The last month is the fourth month in any given interview (the fourth reference month).

1986: All variables that have "4" in it correspond to the last reference month.

1991: REFMTH=4.

1996: SREFMON=4.

labor force status

These are recoded variables concerning labor force status for a given month. If they are equal to 1, 2, 3, 4 or 5, then the person is working. If the variable equals 6 or 7 then the person is looking. When the variable equals 8, the person has not looked or worked during the month. A possible recode for labor force status for a person is: 0 if not in the labor force, 1 if working and 2 if looking.

1986: ESR_4

1991: ESR

1996: RMESR

looking

These are recoded variables concerning labor force status for a given month. If they are equal to 6 or 7, then the person has not worked during the month but is looking for work or on layoff at some point during the month.

1986: ESR_4

1991: ESR

1996: RMESR

modified family ID

This sequence of numbers gives a unique identifier for families when it is important to distinguish between primary family and related subfamilies. If a person belongs to a related subfamily, the subfamily sequence number replaces the family sequence number. Otherwise, the family ID is the same as the "[inclusive family ID](#)".

1986: If S4_NUMBR is greater than zero, then the family ID is:
SS_ID H4_ADDID S4_NUMBR.

1991: If SID is greater than zero, then the family ID is:
SS_ID ADDID SID.

1996: If RSID is greater than zero, then the family ID is:
SSUID SHHADID RSID.

parent

These are the person numbers of the parents. When using these numbers to construct a new “family” ID, remember that the time the parent entered the household may be different from the time the child entered the household (causing the child and the parent to have different ENTRY variables). Also, remember that the family might have changed composition, causing a change in the PNPT variable. This approach may be difficult to use successfully.

1986: PNPT_4

1991: PNPT

1996: EPNMOM, EPNDAD

past participation in food stamps

These variables indicate when a person first received food stamps (this is the month variable; there is another variable for the year). If the person received food stamps in the past, these variables will be greater than zero.

1986: TM8062

1991: TM8062

1996: EFSSTRMN

person ID

These variables make a unique person number that never changes.

1986: SS_ID PP_ENTRY PP_PNUM

1991: Core: SUID ENTRY PNUM
Topical: ID ENTRY PNUM

1996: SSUID EPPPNUM

poverty

A person is in poverty when a family’s income (Census definition) falls below the poverty line.

1986: If F4TOTINC less than F4_POV.

1991: If FTOTINC less than FPOV.

1996: If TFTOTIN less than RFPOV.

reference month

The reference month is the month that the interview is covering. In each interview, SIPP covers the previous 4 months. The variables that state which month the data correspond to are listed in the [last month of the interview](#) variable.

rotation groups

The rotation group variable indicates which group the household belonged to (1–4)

1986: SU_ROT
1991: ROT
1996: SROTATION

sex

If the variable equals 2, then the sex is female.

1986: SEX
1991: SEX
1996: ESEX

working

These are recoded variables concerning labor force status for a given month. If they are equal to 1, 2, 3, 4, or 5, then the person worked at some point during the month.

1986: ESR_4
1991: ESR
1996: RMESR